# *Intelligent Systems Technologies to Assist in Utilization of Earth Observation Data*


- H. (Rama) Ramapriyan*, G. McConaughy*, S. Morse**, D. Isaac***
- * NASA Goddard Space Flight Center, Greenbelt, MD
- ** SoSA Corporation, Chantilly, VA
- *** Business Performance Systems, Bethesda, MD



- Earth Observing Systems IX, SPIE Meeting

- August 6, 2004



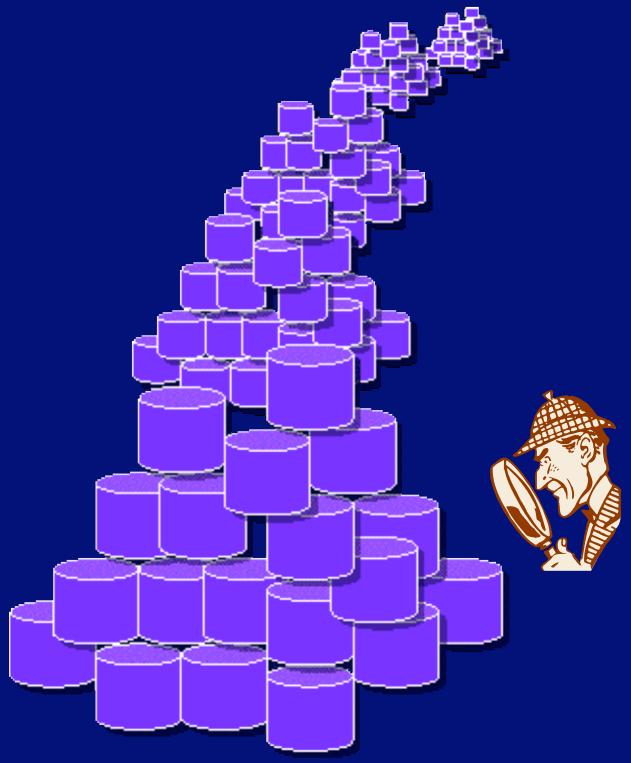
- Rama.Ramapriyan@nasa.gov

# Outline

- The Challenge: Succeeding in a “data rich” environment
- The Opportunity: Knowledge building algorithms
- The Concept: Adding intelligence to data archives
- Results to Date: Assessment of laboratory demonstrations
- Next Steps: Operational proof-of-concept

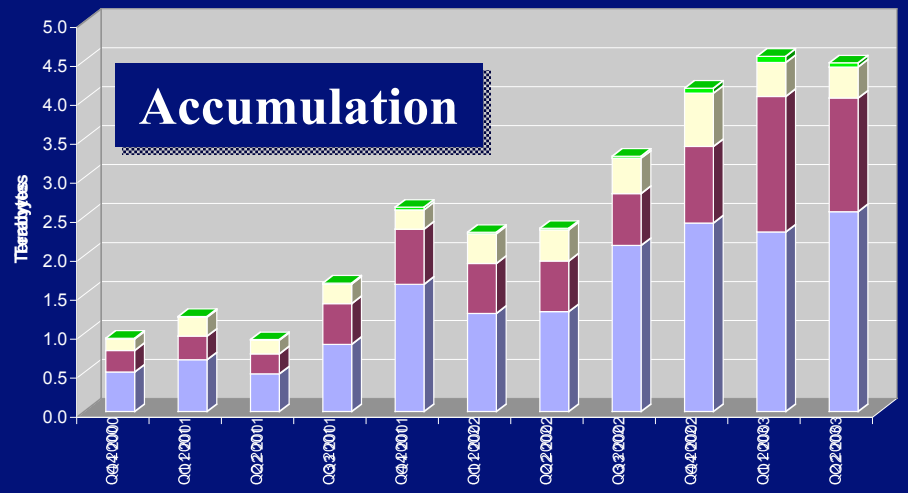
# The Challenge: Succeeding in a Data Rich Environment

- Large and growing data collections from the Earth Observing System
  - 3.4 petabytes of data
  - 48 million files
  - 3.5 terabytes/day accumulation
- Distributed, heterogeneous data systems
  - ~70 data centers
  - Complex “value chains”
- Broad & diverse user community
  - Research, applications, education
- Limited human capacity to examine large volumes of data
  - Users need information, not just data

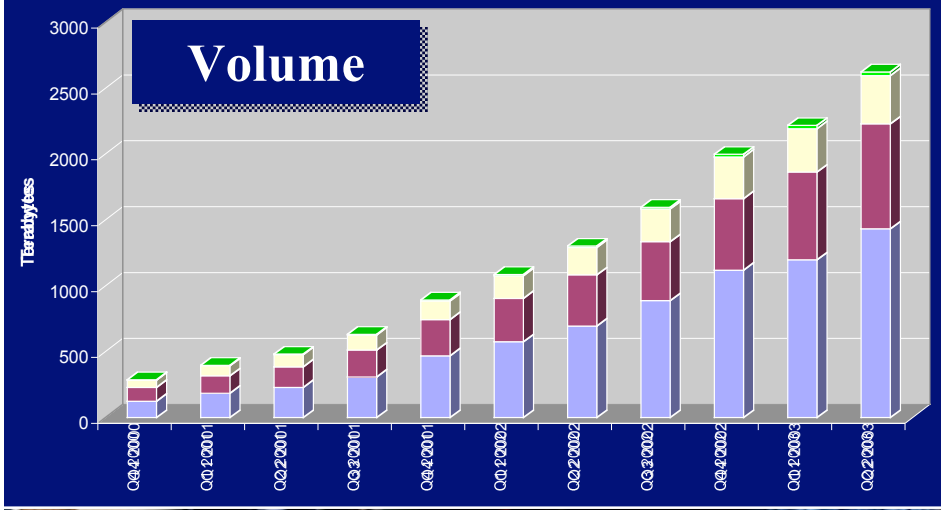


# Large & Growing Earth Science Data Holdings

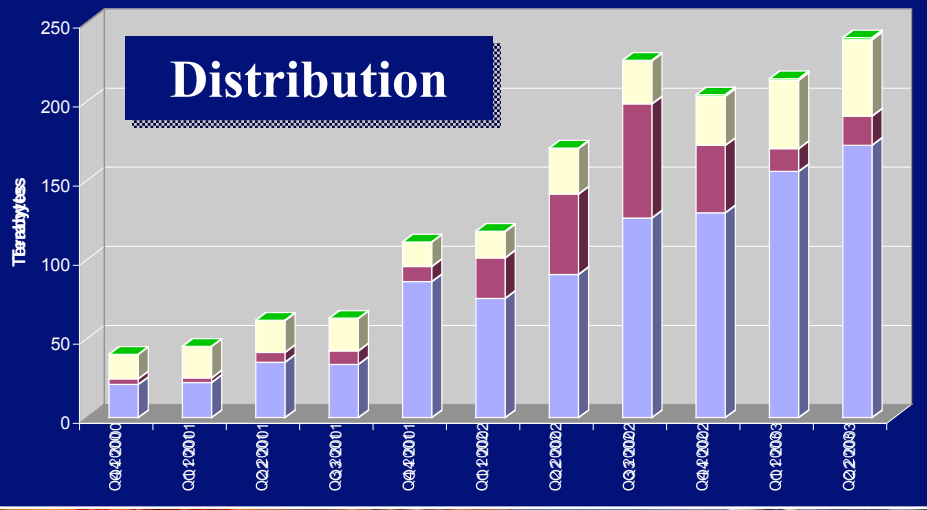
Daily Archive Insert Rate




Cumulative Archive Volume




Distribution Volume




# Widely Distributed and Heterogeneous Data

- Over 70 NASA funded Earth science “data centers” across the US, plus interagency and international partners
- Trend is for further distribution
  - Recent cooperative agreements add to the network of PI systems (REASoN)


# Large & Diverse User Community

- 2.1 million distinct users of EOSDIS in FY2003
  - 228,000 placed data product orders
    - 17,000 placed orders via order system
    - Remainder used on-line download
- Diverse user population requires systems that adapt to different needs
  - Users from commercial, government, educational, international organizations
  - 78% “first time” users in FY2003
- 29 million data products delivered
  - 16% from EOS (L7, Terra, Aqua, SAGE III, ACRIMSat, SORCE)
  - 84% from other Earth science data products (including CERES, Topex, Jason, QuikScat, SeaWinds, Pathfinder data sets, etc.)

# The Challenge: Data Utilization Issues

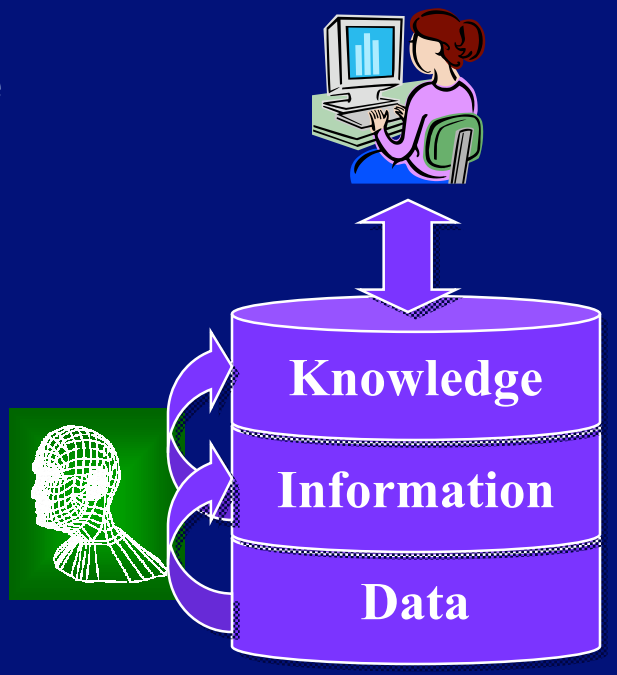
- Timeliness
  - New applications require near-real-time data delivery
  - Human-based data quality assessment can take weeks or longer
- Access
  - Users need more assistance in locating relevant data in large archives
  - Content-based metadata and indexes could help
- Understandability
  - Users need a concise description of the salient characteristics of data
  - But, current data systems are generally oblivious to the content
- Readiness for Use
  - Users want information, not just data
  - Need to move up the data → information → knowledge chain
- Responsiveness
  - Systems should be aware of user needs and adapt to them

# The Opportunity

- Data mining algorithms
  - Induction of general characteristics, relationships, & patterns from specific data
  - Successfully moved from labs to industrial use
- Intelligent data understanding
  - Research sponsored by NASA's Intelligent Systems Project
  - 22 research projects exploring a variety of algorithms applied to a variety of data...including remote sensing data
- Affordable high-performance computing
  - Improvements may make large-scale data mining feasible
  - Grid technologies could also provide needed capacity

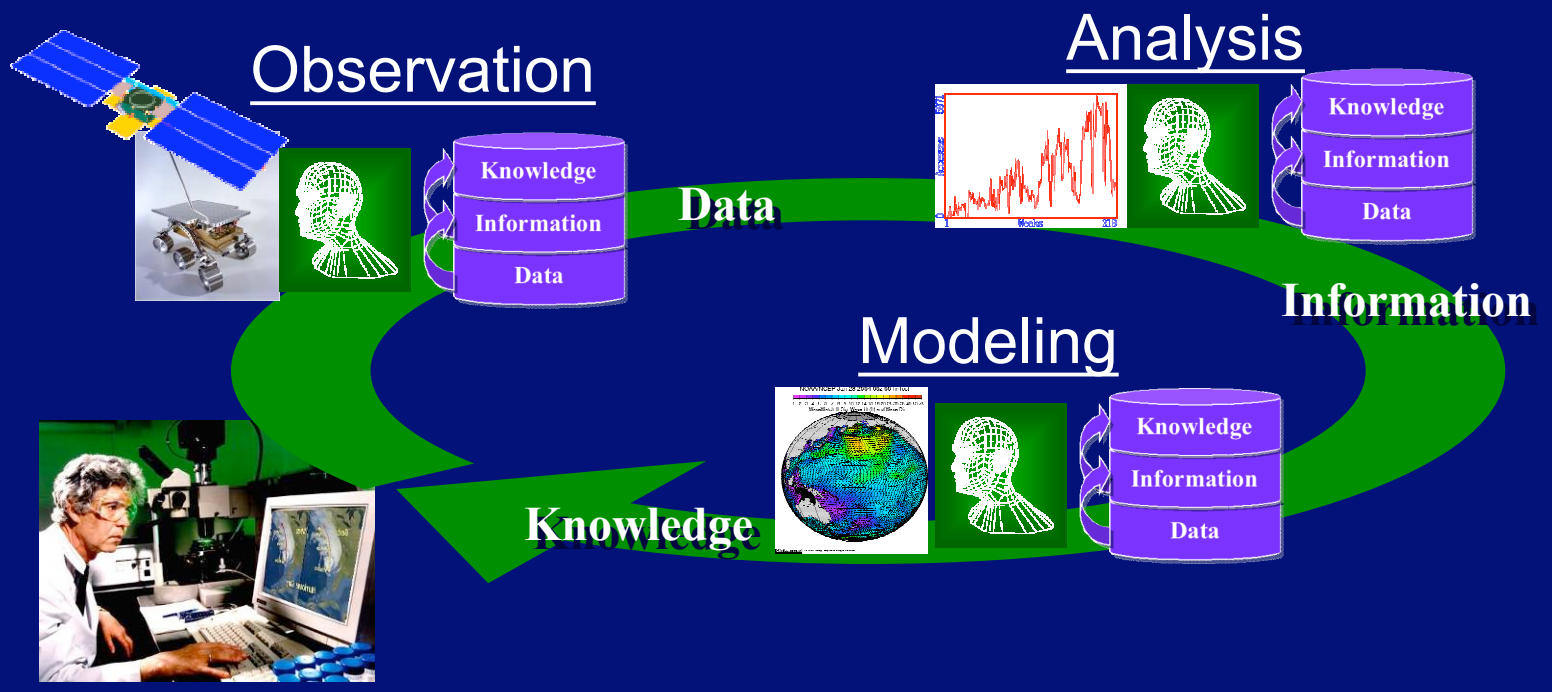
# The Concept

- Intelligent Archives
  - Archive is aware of its own data content and usage
  - Archive can extract new information from data holdings
- Knowledge Building Systems
  - Directly support building knowledge from data and information
  - Incorporates intelligent archives to extract information & knowledge
  - Includes feedback loops to improve adaptation to user needs and external events
  - Includes coordination between intelligent archives and intelligent sensors
  - Highly distributed and collaborative




# Intelligent Archives in the Context of Knowledge Building Systems (IA-KBS)

- Data archives exist throughout the information value chain
- Intelligence with feedback loops makes systems more effective
- Distributed intelligent components collaborate to achieve user goals




# Intelligent Archives in the Context of Knowledge Building Systems (cont'd)

**ENTERPRISE**
(earth and space sciences)

Supported by
**DISTRIBUTED INTELLIGENT SYSTEMS**

Integrated through
**DISTRIBUTED INFRASTRUCTURES**

Transformation Loop

Knowledge

Information

Data

Observation

Applications

**Intelligent Functions & Services:**

- **Data Understanding**
- **Data Management**
- **Data Persistence and Preservation Management**

Processing Systems

Intelligent Sensors

Infrastructures of Physical and Virtual computing resources, services, communications

# IA-KBS Potential Capabilities

- Virtual product generation
  - Dynamically assemble an information product specific to the user's need from relevant data
  - Intelligence needed to understand data relationships relative to an information "goal" and anticipate user requests
- Significant event detection
  - Automatically learn "normal" data streams and identify exceptions
  - Intelligent archive can focus attention on interesting data subsets
- Automated data quality assessment
  - Automatically identify anomalies in the data stream
  - Relieves human burden and enables rapid quality assessments

# IA-KBS Potential Capabilities (cont'd)

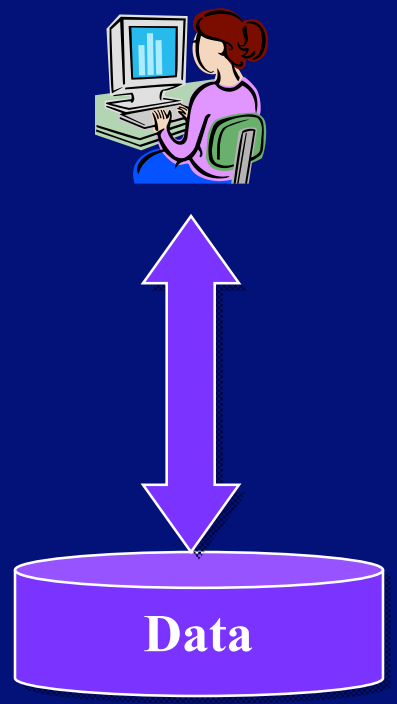
- Large-scale data mining
  - Continuously mine archived data searching for hidden relationships and patterns
  - Enables archive to suggest models for human evaluation
- Dynamic feedback loop
  - Acting on information discovered, such as a significant event
  - Enables archive to adapt to events and anticipate user needs
- Data discovery and efficient requesting
  - Identifying new data sources and information collaborators, and using available resources judiciously
  - Enables archive to reach farther than its own holdings

# IA-KBS – Relevant Technologies

- Distributed system architectures
  - Especially, Grid technologies
- Intelligent data understanding algorithms
  - Fern & Brodley: understanding high-dimensionality data using clustering, re-projection, cluster ensembles
  - Kumar et al: discovering climate indices using clustering on time-series data
  - Teng: identifying and removing anomalies to improve classifier performance
  - Kargupta: extending data mining algorithms to distributed architectures
  - Smelyanskiy: Bayesian inference of non-linear dynamical model parameters
  - Nemani & Golden: dynamic assembly of data and operators to satisfy a user's information goal
  - LeMoigne: sub-pixel accurate image registration for data fusion

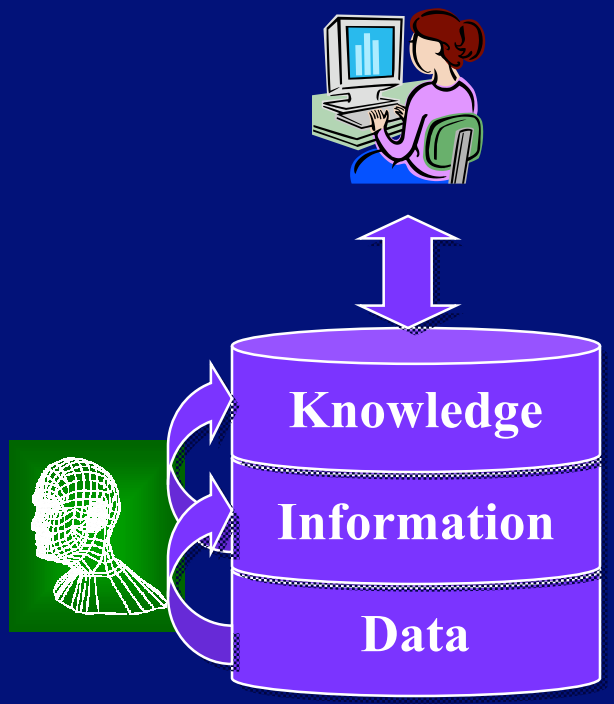
# IA-KBS Vision: Before

- User: “Retrieve estimated precipitation for Summer 2005 for mid-west US”
  - System: “No product matched the search criteria”
- User: “Retrieve precipitation for 2000-2004 for mid-west US”
  - System: “115,285 granules found...hygrometer profile, radar rainfall rate, combined rainfall profile, monthly 5x5 degree spaceborne radar rainfall, ...
- User: “?”




# IA-KBS Vision: After

- User: “Retrieve estimated precipitation for Summer 2005 for mid-west US”
  - System:
    - Precipitation is correlated to sea surface temperature at region labeled “El Nino Southern Oscillation”
    - Estimated precipitation is calculated from stored model
    - Precipitation data has been pre-staged based on prior queries indicating user interests in drought-related data
    - Sea surface temperature anomaly information has been retrieved from collaborating distributed archive
  - System: “Estimated precipitation is 1.2” below average for the specified area and time.”




# Conclusions

- Intelligent archives can improve the utility of data
  - Improved timeliness, ease of access, understandability, readiness for use, and responsiveness
- Intelligent archives can enable a variety of needed capabilities
  - Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting.
- Promising data mining algorithms have been identified and applied to remote sensing data in a laboratory environment
- Next step is to demonstrate utility and scalability in an operational environment